Evaluación de la competencia léxica de los LLM: Un estudio comparativo con el corpus CAES basado en el MCER

Evaluating LLM lexical competence: A comparative CEFR-based study with the CAES corpus

M. Victoria Cantero Romero,[1] Isabel Cabrera de Castro,[2] Jaime Collado Montañez,[2] Salud María Jiménez-Zafra[2]
[1]Departamento de Filología Española, SINAI, CEATIC, Universidad de Jaén, España
[2]Departamento de Informática, SINAI, CEATIC, Universidad de Jaén, España
{vcantero, iccastro, jcollado, sjzafra}@ujaen.es

Resumen: Los modelos de lenguaje de gran tamaño (LLM) han adquirido un papel relevante en el ámbito educativo, especialmente en la enseñanza de español como lengua extranjera, donde se emplean para la generación de contenidos, la corrección automática y el desarrollo de sistemas conversacionales. No obstante, resulta necesario evaluar la adecuación de su producción léxica a los niveles de competencia definidos por el Marco Común Europeo de Referencia. En este trabajo se analiza la complejidad léxica de seis LLM mediante una comparación con el corpus CAES. Para ello, se generaron textos a partir de tareas equivalentes a las realizadas por estudiantes, sin especificar el nivel de dificultad. Asimismo, se construyó un lexicón anotado por niveles (A1–C1) a partir del corpus de referencia. Los resultados indican que, aunque los LLM reproducen tendencias similares a las de los aprendices, estos muestran un uso más amplio de vocabulario especializado en niveles avanzados.
Palabras clave: Modelos de lenguaje de gran tamaño, nivelación léxica, complejidad léxica, español como lengua extranjera.

Abstract: Large language models (LLM) have come to play a significant role in the field of education, particularly in the teaching of Spanish as a foreign language, where they are used for content generation, automatic correction and the development of conversational systems. However, it is necessary to assess the suitability of their lexical output for the proficiency levels defined by the Common European Framework of Reference. This study analyses the lexical complexity of six LLM by comparing them with the CAES corpus. To this end, texts were generated based on tasks equivalent to those performed by students, without specifying the level of difficulty. Furthermore, a lexicon annotated by levels (A1–C1) was constructed from the reference corpus. The results indicate that, although LLM reproduce trends similar to those of learners, they demonstrate a broader use of specialised vocabulary at advanced levels.
Keywords: Large language models, lexical levelling, lexical complexity, Spanish as a foreing language

1 Introducción

La irrupción de los modelos de lenguaje de gran tamaño (LLM) en el ámbito de la enseñanza del Español como Lengua Extranjera (ELE) ha transformado las posibilidades de generación de contenido y evaluación automatizada. Sin embargo, la integración efectiva de la Inteligencia Artificial (IA) en el aula no solo requiere fluidez generativa, sino una precisión pedagógica que se ajuste estrictamente a los estándares internacionales de competencia lingüística. En este sentido, la evaluación de la calidad léxica constituye uno de los pilares fundamentales para determinar el nivel de dominio de un aprendiz.

El Marco Común Europeo de Referencia (MCER) y el Plan Curricular del Instituto Cervantes (PCIC) definen, respectivamente,

los niveles de competencia internacional y los contenidos específicos para la enseñanza del español, constituyendo un marco de referencia sobre la progresión del vocabulario, desde los niveles iniciales de acceso (A1) hasta el dominio experto (C1/C2). A pesar de la aparente fluidez de los modelos actuales, hasta donde sabemos, no existen estudios acerca de si el vocabulario que emplean los LLM en tareas específicas se alinea realmente con el léxico que un estudiante de un determinado nivel (A1, A2, B1, B2, C1 o C2) utiliza habitualmente, o con lo que los estándares curriculares exigen.

El objetivo principal de este trabajo es determinar el nivel léxico real que poseen diversos LLM en comparación con las producciones de estudiantes de ELE y los descriptores del MCER. Para ello, se han evaluado modelos de código abierto como Gemma, Llama 3.3, DeepSeek y GPT-OSS, y modelos cerrados como GPT-5.4 y Gemini Pro. Como fuente empírica de producciones reales de estudiantes, se ha usado el Corpus de Aprendices de Español (CAES), considerado el corpus de referencia en la nivelación de estudiantes de ELE. A partir de este corpus, se ha generado un lexicón nivelado que permite categorizar el vocabulario según los niveles de competencia del MCER. La metodología consiste en solicitar a los modelos la generación de textos para las mismas tareas de expresión escrita recogidas en el CAES, evaluando posteriormente estas producciones mediante el lexicón extraído. Este diseño experimental permite observar si el uso "cotidiano" de la IA emula el comportamiento humano en tareas donde la complejidad léxica va aumentando gradualmente.

El resto del artículo se estructura de la siguiente manera. En la Sección 2, se revisa el estado de la cuestión sobre la competencia léxica en ELE y el marco de referencia, las medidas de complejidad léxica, y los modelos de lenguaje en la enseñanza del español. En la Sección 3, se describe la metodología de construcción del lexicón a partir del corpus CAES, detallando el procesamiento automático de las muestras, los criterios de selección de palabras con carga semántica y el procedimiento de filtrado progresivo empleado para asignar el vocabulario a los distintos niveles del MCER. La Sección 4 detalla el diseño experimental, especificando los criterios de selección de los modelos, la configuración de las estrategias de *prompting* y la réplica de las tareas de escritura utilizadas para la evaluación. En la Sección 5 se presentan y analizan los resultados obtenidos contrastando el perfil léxico de los LLM con la producción real de los aprendices, discutiendo su grado de ajuste a la progresión del MCER. Por último, la Sección 6 sintetiza las conclusiones principales del estudio, reflexionando sobre las implicaciones pedagógicas, y plantea líneas de trabajo futuro.

2 Estado de la cuestión

La evaluación de la competencia lingüística en el aprendizaje de segundas lenguas ha evolucionado significativamente, integrando herramientas computacionales y modelos estadísticos para medir la calidad de las producciones. En el contexto de ELE, el análisis del léxico se ha convertido en un componente esencial para determinar el nivel de dominio de los estudiantes y, recientemente, para evaluar las capacidades de los LLM.

2.1 La competencia léxica en ELE y el marco de referencia

La competencia léxica se define como el conocimiento y la capacidad para utilizar el vocabulario de una lengua de la misma manera que los hablantes nativos, incluyendo el dominio de las familias de palabras y las colocaciones comunes (López-Mezquita Molina, 2007). En el ámbito europeo, el MCER y PCIC establecen las directrices para la progresión del aprendizaje, definiendo niveles que van desde el A1 (acceso) hasta el C2 (maestría) (Consejo de Europa, 2020).

La adecuación del léxico a estos niveles es fundamental para asegurar una progresión efectiva del estudiante. Para ello, el PCIC detalla nociones específicas y descriptores de riqueza de vocabulario para cada uno de ellos. Por ejemplo, en el nivel B1 (umbral), el estudiante debe poseer suficiente vocabulario para expresarse sobre temas cotidianos, mientras que en el B2 (avanzado) debe ser capaz de variar la formulación para evitar repeticiones frecuentes (Instituto Cervantes, 2006).

2.2 Medidas de complejidad léxica

En la lingüística cuantitativa, el concepto de riqueza léxica o complejidad léxica se utiliza para medir la cantidad y calidad del vocabulario empleado en una muestra de lengua. Es-

te constructo es multidimensional y se compone de al menos tres aspectos principales:

- Diversidad léxica: Se refiere a la variedad de palabras únicas (*types*) frente al total de palabras (*tokens*) en un texto. La medida convencional es la *Type-Token Ratio (TTR)*, aunque es sensible a la longitud del texto.
- Sofisticación léxica: Caracteriza el uso de palabras poco frecuentes o avanzadas que indican un nivel de conocimiento superior. Generalmente se mide contrastando el léxico con listas de frecuencia de un corpus de referencia.
- Densidad léxica: Es la proporción de palabras con contenido léxico (sustantivos, verbos, adjetivos, adverbios) frente a las palabras funcionales.

Para automatizar estos cálculos, se han desarrollado herramientas como *Lexical Frequency Profile* (LFP) (Laufer y Nation, 1995), que divide las palabras en franjas de frecuencia (1K, 2K, etc.), y TAALES (*Tool for the Automatic Analysis of Lexical Sophistication*) (Kyle y Berger, 2017), que permite obtener cientos de índices relacionados con la sofisticación.

2.3 Modelos de lenguaje en la enseñanza del español

Los LLM han demostrado una gran capacidad para generar textos coherentes y contextualizados, lo que los convierte en herramientas valiosas para la creación de materiales didácticos y la evaluación del progreso estudiantil (Mullins et al., 2024).

Sin embargo, estudios recientes indican que los textos que estos generan para niveles específicos de ELE no siempre se ajustan estrictamente a los estándares del PCIC. En niveles básicos como A1 y A2, modelos como GPT-3.5 y LLaMA-3.1 tienden a producir un vocabulario con un nivel de complejidad superior al solicitado. Por el contrario, en niveles intermedios como B1 y B2, se ha observado que GPT-3.5 muestra una mayor coherencia y adecuación léxica que LLaMA-3.1, el cual presenta fluctuaciones notables al incluir términos de niveles inferiores y superiores de manera errática (Cantero Romero, 2025).

A pesar de su popularidad, el conocimiento real que tienen los LLM de los idiomas distintos al inglés no ha sido tan estudiado. Conde et al. (2024) demostraron que dos tercios de los modelos evaluados no logran producir significados válidos para más de la mitad de las palabras de un diccionario de referencia, y que la capacidad para utilizar palabras correctamente en frases relacionadas es aún menor, situándose por debajo del 25 % en la mayoría de los modelos.

En definitiva, aunque existen grandes esfuerzos para adaptar modelos al español (Gonzalez-Agirre et al., 2025), los LLM actuales presentan un conocimiento léxico limitado y una adecuación irregular a los marcos de competencia lingüística humana, lo que subraya la necesidad de revisiones manuales y del desarrollo de lexicones nivelados específicos para su evaluación.

2.4 Evaluación de la generación de modelos de lenguaje

Investigaciones recientes han analizado exhaustivamente la calidad de textos generados artificialmente en inglés mediante parámetros léxicos y de cohesión, encontrando que la similitud entre los textos generados por IA y los escritos por humanos es extremadamente alta, oscilando en algunos índices cerca del 100 % (Skowron y Baczkowska, 2023). No obstante, se ha observado que no existe un modelo único que supere a los demás en todos los indicadores lingüísticos.

Un aspecto crítico en la evaluación de estos modelos es su diversidad léxica, entendida como la variedad de vocabulario único empleado. Estudios comparativos (Reviriego et al., 2024) indican una evolución significativa: mientras que versiones iniciales como GPT-3.5 suelen mostrar una diversidad léxica menor a la de los humanos en tareas de parafraseo y respuesta a preguntas, modelos más recientes como GPT-4 han logrado igualar, e incluso superar en ciertos contextos, la riqueza de vocabulario de estos.

Además, en el ámbito educativo, se ha demostrado que los ensayos argumentativos generados por GPT-4 reciben calificaciones de calidad superiores a las de los estudiantes humanos en criterios como estructura lógica, complejidad del lenguaje y riqueza de vocabulario (Herbold et al., 2023). No obstante, se ha identificado que la IA tiende a utilizar un estilo más científico y complejo, caracterizado por una mayor nominalización y complejidad

sintáctica.

Más recientemente, se han publicado trabajos centrados en la diversidad léxica y lingüística de los modelos de lenguaje en lengua inglesa (Kendro, Maloney, y Jarvis, 2026; Guo, Shang, y Clavel, 2025), que han abordado de forma específica esta cuestión. Kendro, Maloney, y Jarvis (2026) muestran que la similitud entre la diversidad léxica de los LLM y la escritura humana depende en gran medida del tipo de tarea, observándose que los modelos pueden igualar o incluso superar la riqueza léxica humana en tareas académicas y argumentativas, mientras que presentan un comportamiento menos diverso en tareas creativas. Por su parte, Guo, Shang, y Clavel (2025) proponen una evaluación multidimensional de la diversidad lingüística de los LLM y concluyen que, pese a su elevada fluidez, estos presentan patrones léxicos diferenciados respecto a la producción humana. Estos trabajos ponen de manifiesto la importancia de analizar el comportamiento léxico de los modelos en contextos de uso específicos, aspecto que motiva el enfoque adoptado en este trabajo para el caso del español como lengua extranjera.

3 Metodología

En esta sección se presenta la metodología seguida en el estudio. Se describe, en primer lugar, el corpus de referencia utilizado y, posteriormente, el proceso de construcción del lexicón nivelado.

3.1 Corpus

Para llevar a cabo la nivelación del léxico utilizado por distintos LLM, se ha construido un lexicón a partir del Corpus de Aprendices de Español CAES (Universidad de Santiago de Compostela, 2022).

El corpus CAES fue recopilado por la Universidad de Santiago de Compostela y financiado por el Instituto Cervantes. Los datos fueron recogidos entre octubre de 2011 y diciembre de 2020 en centros educativos, principalmente universitarios, de distintos países como España, Estados Unidos, Brasil, Egipto, Irlanda o Portugal. Los participantes procedían de once lenguas maternas diferentes (inglés, chino mandarín, portugués, árabe, ruso, alemán, francés, griego, italiano, japonés y polaco).

En este trabajo se ha utilizado la versión 2.1 del corpus (marzo de 2022), que amplía la primera recopilación e incluye 2544 participantes, frente a los 1423 de la versión inicial.

El corpus contiene textos escritos por aprendices de español clasificados según los niveles del MCER desde A1 hasta C1. En la Tabla 1 podemos ver una muestra de las diferentes tareas. Las tareas corresponden a distintos géneros discursivos acordes con las funciones comunicativas descritas en el PCIC, como correos electrónicos, biografías, cartas, narraciones, solicitudes o reseñas. En los niveles iniciales (A1 y A2) se recogen aproximadamente 2000 textos por nivel, mientras que en los niveles avanzados (B1, B2 y C1) el número de muestras disminuye debido al menor número de tareas.

La anotación del corpus fue realizada por especialistas en enseñanza de español como lengua extranjera de la Universidad de Santiago de Compostela. Cada texto fue clasificado según los niveles del MCER siguiendo los criterios establecidos en el PCIC y las pautas definidas en el propio proyecto (Palacios Martínez, Barcala Rodríguez, y Rojo, 2019). Para garantizar la fiabilidad de la clasificación, los textos fueron evaluados de forma independiente por varios anotadores y, posteriormente, revisados hasta alcanzar consenso.

3.2 Lexicón

A partir del corpus CAES se ha elaborado un lexicón nivelado según los niveles del MCER. Para ello, los textos fueron procesados automáticamente mediante la herramienta de PoS tagging incluida en SpaCy (Honnibal y Montani, 2017).

En el proceso de extracción se seleccionaron únicamente palabras con carga semántica (sustantivos, verbos, adjetivos y adverbios). Esta decisión se fundamenta en que las palabras léxicas constituyen el principal indicador del repertorio léxico de un hablante, mientras que las palabras funcionales presentan menor variación entre niveles de competencia (Laufer y Nation, 1995).

Tras la identificación de las categorías gramaticales, se calculó la frecuencia de aparición de cada palabra en los textos correspondientes a cada nivel del MCER (A1, A2, B1, B2 y C1). Para determinar el vocabulario característico de cada nivel, se aplicó un proceso de filtrado progresivo: las palabras presentes en niveles inferiores se eliminaron de los niveles superiores. De este modo, en A2 se

Nivel	Tarea	Muestras
A1	Correo electrónico cambio de trabajo	728
	Correo electrónico familia	703
	Nota llegar tarde	705
A2	Biografía persona que admiras	673
	Reserva habitación de hotel	603
	Postal de vacaciones	701
B1	Carta a un amigo	528
	Historia graciosa	382
	Reclamación compañía aérea	454
B2	Solicitud admisión	375
	Fumar en lugares públicos	356
C1	Reseña película	184
	Reclamación compañía gas	169

Tabla 1: Resumen de tareas y muestras del corpus CAES empleadas en el estudio.

excluyeron las palabras ya presentes en A1, en B1 las de A1 y A2, y así sucesivamente, lo que permite identificar el vocabulario predominante en cada nivel.

No se realizó lematización del vocabulario con el fin de preservar la información morfológica de las formas flexionadas. En particular, los morfemas verbales que indican tiempo o modo constituyen indicadores relevantes del nivel de competencia lingüística, ya que el uso de determinados tiempos verbales suele asociarse a niveles más avanzados.

Asimismo, no se tuvo en cuenta la polisemia de las palabras extraídas, dado que el análisis se realizó fuera de contexto. El tratamiento de los distintos sentidos léxicos se plantea como una posible línea de trabajo futuro. El análisis se centra exclusivamente en el vocabulario utilizado en tareas de expresión escrita, es decir, en el vocabulario activo presente en los textos producidos. En consecuencia, el estudio no pretende estimar el vocabulario potencial que los modelos puedan conocer, sino únicamente el léxico que efectivamente utilizan en sus producciones.

Dado que en los niveles más avanzados del MCER la progresión léxica es fundamentalmente cualitativa, centrada en el dominio flexible del repertorio previamente adquirido, el análisis se ha limitado a los niveles A1–C1. En el Plan Curricular del Instituto Cervantes, el nivel C2 se caracteriza principalmente por un uso preciso y matizado del léxico disponible en niveles anteriores.

4 Experimentación

En esta sección se describe la configuración experimental, incluyendo los modelos evaluados, las tareas y los prompts utilizados, así como el procedimiento de análisis de las salidas generadas por los modelos.

Modelo	Parámetros
Gemma 3 27B IT	27B
gpt-oss-120b	116.8B
Llama 3.3 70B Instruct	70B
DeepSeek-V3	671B
Gemini Pro	-
GPT-5.4	-

Tabla 2: Modelos evaluados y número de parámetros.

4.1 Modelos evaluados

Para la experimentación hemos seleccionado un grupo diverso de arquitecturas, que incluye tanto opciones de pesos abiertos disponibles para su despliegue local, como sistemas propietarios cerrados que conforman el estado de la cuestión en este campo. En concreto, las alternativas abiertas evaluadas fueron Gemma 3 (Google, 2025), GPT-OSS (OpenAI, 2025), Llama 3.3 (Meta, 2024) y DeepSeek v3.2 (DeepSeek-AI, 2024), mientras que, entre los modelos cerrados, se analizaron Gemini 3 Pro (Google, 2026) y GPT-5.4 (OpenAI, 2026). Esta selección permite observar posibles diferencias según la naturaleza de su desarrollo y comparar el comportamiento de las soluciones abiertas frente a sus equivalentes comerciales. La Tabla 2 resume los modelos considerados y los principales parámetros empleados durante la generación.

4.2 Prompts y tareas

La evaluación se llevó a cabo mediante tareas de producción escrita basadas en las utilizadas en el corpus CAES. Estas actividades siguen una progresión de complejidad asociada

a los niveles del MCER en la destreza de expresión escrita, por lo que resultan adecuadas para observar si el vocabulario generado por los modelos aumenta también en complejidad a medida que avanzan las tareas. La Tabla 1 recoge las tareas empleadas en el experimento.

Dado que las tareas difieren en temática y género discursivo, el vocabulario producido puede verse condicionado en parte por el contenido solicitado. No obstante, esta variación forma parte del propio diseño de las tareas y refleja distintas situaciones comunicativas asociadas a cada nivel.

La generación se realizó mediante una estrategia de *zero-shot prompting*, en la que se proporcionó únicamente la instrucción correspondiente a cada tarea, sin ejemplos adicionales. Esta decisión responde al objetivo de evaluar el conocimiento léxico propio del modelo sin incorporar información externa a través del *prompt*, evitando así posibles sesgos en la generación.

Durante el diseño experimental se probaron dos formulaciones de *prompt*: una con rol explícito del hablante y otra sin rol. En la primera, se atribuía al modelo la identidad de un estudiante de español correspondiente a un nivel determinado; en la segunda, se presentaba únicamente la instrucción de escritura. Las pruebas preliminares realizadas con Gemma y Gemini no mostraron diferencias relevantes en el tipo de vocabulario generado entre ambas formulaciones. Por este motivo, se optó por utilizar *prompts* sin rol explícito, con el fin de evitar un posible sesgo inducido por la instrucción y observar con mayor fidelidad el comportamiento léxico habitual del modelo. A partir de esta decisión, los *prompts* utilizados en la experimentación adoptaron una formulación directa basada en la consigna de escritura correspondiente a cada tarea. Su estructura general fue la siguiente:

> *Eres un alumno y tienes que [Descripción de la tarea]. Entre [mínimo] y [máximo] palabras.*

Por ejemplo:

> *Eres un alumno y tienes que escribir un correo electrónico a unos compañeros de clase o de trabajo presentándote. Entre 75 y 100 palabras.*

4.3 Salidas generadas

Con el fin de analizar el comportamiento léxico de los modelos de lenguaje, se diseñó un procedimiento de evaluación automatizada basado en el lexicón de referencia extraído del corpus CAES. En primer lugar, se aplicó un etiquetado morfosintáctico (*Part-of-Speech tagging*) a los textos generados por cada modelo mediante la biblioteca spaCy y el modelo `es_core_news_md`, extrayendo únicamente las categorías gramaticales con carga semántica consideradas en este estudio y obteniendo la forma lematizada de cada token. A continuación, cada palabra extraída se contrastó con el lexicón de referencia construido a partir del corpus CAES, en el que cada ítem léxico había sido previamente asociado a un nivel del MCER (A1, A2, B1, B2 o C1).

Para cada texto generado se calculó el porcentaje de palabras pertenecientes a cada nivel léxico, determinando el nivel predominante como aquel con mayor representación porcentual. Este procedimiento permitió comparar el perfil léxico de los textos generados por los modelos con el de los textos producidos por aprendices reales del corpus CAES en tareas equivalentes.

5 Resultados

El presente apartado expone los resultados del análisis léxico obtenido en las tareas correspondientes a los distintos niveles de competencia lingüística del MCER. Para cada nivel (A1–C1), se comparan los índices de distribución léxica generados por diferentes modelos de lenguaje con los valores de referencia del corpus CAES. El objetivo es identificar patrones de adecuación léxica, así como posibles desviaciones en la producción de vocabulario pertenecientes a otros niveles de competencia. Esta comparación permite evaluar la sensibilidad de los modelos al gradiente de dificultad léxica y su capacidad para ajustar la complejidad del vocabulario al nivel requerido.

5.1 Nivel A1

Los índices de palabras correspondientes a las tareas del nivel A1 (Tabla 3) muestran que el léxico predominante se ajusta, en términos generales, a lo esperado para este nivel inicial. No obstante, la mayoría de los modelos tienden a incorporar una proporción mayor de palabras procedentes de otros niveles

Modelo	A1	A2	B1	B2	C1
GPT-OSS-120b	82,3	9,83	3,33	0	0
GPT-5.4	94,77	0,6	0	0	0
Gemma-3-27b-it	91,9	2,4	1,87	0,53	0
Gemini-3-pro	94,57	3,67	0	0	0
Deepseek-v3.2	83,43	2,9	3,83	2,3	0,57
Llama-3.3-70b-instruct	80,3	11,7	6,67	0	0
CAES	99,27	0	0	0	0

Tabla 3: Resultados tareas A1.

Modelo	A1	A2	B1	B2	C1
GPT-OSS-120b	72,07	16,27	2,17	2,2	0,53
GPT-5.4	83	10.73	1.1	0	0.57
Gemma-3-27b-it	73,8	14,13	2,97	2,33	0,97
Gemini-3-pro	81,8	9,77	2,73	0	1,03
Deepseek-v3.2	69,77	16,73	2,03	1,57	1,57
Llama-3.3-70b-instruct	77.77	14.6	0.6	0.6	0.6
CAES	87,47	11,95	0	0	0

Tabla 4: Resultados tareas A2.

de competencia, como ocurre de manera especialmente evidente en el modelo Deepseek-v3.2, que incluye incluso vocabulario categorizado en el nivel C1.

En la comparativa entre modelos abiertos y cerrados, se observa que los modelos cerrados (GPT-5.4 y Gemini-3-pro) presentan resultados más próximos a los valores del corpus CAES, siendo los que utilizan en menor medida léxico perteneciente a otros niveles.

5.2 Nivel A2

En relación con las tareas del nivel A2 (Tabla 4), se aprecia una disminución del léxico de nivel A1 y un incremento del léxico A2, en consonancia con la progresión esperada en la complejidad lingüística. Sin embargo, a diferencia de lo observado en el corpus CAES, todos los modelos continúan incorporando vocabulario de niveles superiores, alcanzando en algunos casos el nivel C1.

Los modelos abiertos superan los porcentajes de léxico A2 registrados en el CAES, mientras que GPT-5.4 y Gemini-3-pro se sitúan por debajo de dicho valor. De este modo, los resultados evidencian una doble tendencia: Deepseek-v3.2 (16,73 %) y GPT-OSS-120b (16,27 %) tienden a sobreproducir léxico A2, mientras que Gemini-3-pro (9,77 %) y GPT-5.4 (10,73 %) lo infraproducen, aunque con valores relativamente cercanos al porcentaje de referencia del corpus CAES (11,95 %).

5.3 Nivel B1

Los datos asociados al nivel B1 (Tabla 5) revelan una diferencia notable en la presencia de léxico de los niveles A1 y A2 entre los modelos y el corpus CAES, ya que este último tiende a reducir en mayor medida la cantidad de vocabulario categorizado en dichos niveles básicos. En lo que respecta al propio nivel B1, únicamente dos modelos superan el porcentaje del CAES: GPT-OSS-120b y Deepseek-v3.2, coincidiendo además con los únicos modelos que no generaron palabras clasificadas en el nivel C1.

5.4 Nivel B2

Para el nivel B2 (Tabla 6), los resultados indican que los modelos superan el porcentaje de léxico B2 registrado en el corpus CAES, siendo Gemini (10,85 %) y Deepseek (11,5 %) los que presentan la mayor carga léxica de este nivel. Ambos modelos son, asimismo, aquellos cuyo índice de léxico de niveles A1 y A2 resulta más bajo en comparación con el resto de los modelos analizados.

5.5 Nivel C1

Por último, los resultados relativos al nivel C1 (Tabla 7) muestran que el corpus de referencia establece un valor esperado del 6,81 % de léxico C1 (un porcentaje relativamente elevado que refleja la presencia genuina de vocabulario avanzado en textos de este nivel); sin embargo, ningún modelo logra aproximarse a dicho valor. La brecha oscila entre -1,51 puntos porcentuales en el caso de GPT-OSS-120b (5,30 %) y - 4,76 puntos porcentuales en GPT-5.4 (2,05 %), lo que representa un déficit de entre el 22 % y el 70 % respecto al léxico avanzado esperado.

Modelo	A1	A2	B1	B2	C1
GPT-OSS-120b	71,83	9,3	9,17	3,07	0,63
GPT-5.4	76,93	6,7	7,5	1,77	0
Gemma-3-27b-it	72,33	10,13	7,83	0,63	0,63
Gemini-3-pro	68,3	10,2	7,57	3,17	0,77
Deepseek-v3.2	67,5	11,4	9,43	1,7	1,7
Llama-3.3-70b-instruct	76,53	7,97	7,77	1,23	0
CAES	83,24	6,62	8,43	0	0

Tabla 5: Resultados tareas B1.

Modelo	A1	A2	B1	B2	C1
GPT-OSS-120b	58,9	15,45	9,65	9,4	1,9
GPT-5.4	66	14,7	8,15	7,4	0,3
Gemma-3-27b-it	64,9	13,5	8,85	8	0,55
Gemini-3-pro	54,95	13,45	9,4	10,85	1,45
Deepseek-v3.2	55,8	14,35	7,75	11,5	1,35
Llama-3.3-70b-instruct	62,65	17,8	7,2	9,5	0,35
CAES	79,27	9,18	3,36	7,23	0

Tabla 6: Resultados tareas B2.

Modelo	A1	A2	B1	B2	C1
GPT-OSS-120b	52,8	14,85	9,6	8,2	5,3
GPT-5.4	60,2	19,35	9,65	4,1	2,05
Gemma-3-27b-it	57,5	17,4	7	6,4	3,5
Gemini-3-pro	56,4	15,95	7,55	4,45	4,35
Deepseek-v3.2	55,7	17,45	7,15	4,05	2,65
Llama-3.3-70b-instruct	66,75	14,3	6,95	4,35	3,35
CAES	77,29	9,38	4,26	1,19	6,81

Tabla 7: Resultados tareas C1.

Desde la perspectiva del análisis del corpus, se identifica un patrón característico de curva en J invertida (Figura 1), en el que se produce un repunte en la proporción de léxico en el nivel C1, con un incremento esperado de 5,62 puntos porcentuales. Este comportamiento no es reproducido por ninguno de los modelos analizados, cuyos resultados muestran una disminución monotónica sin registrar dicho repunte en los niveles avanzados. Este hallazgo sugiere que los modelos presentan dificultades para detectar la mayor densidad relativa de vocabulario de nivel C1 en las tareas correspondientes, posiblemente debido a una tendencia a clasificar el léxico avanzado como perteneciente a los niveles B1 o B2.

6 Conclusiones y líneas de trabajo futuro

El análisis del vocabulario empleado por los modelos de lenguaje pone de manifiesto una clara concentración en los niveles básicos de competencia lingüística, especialmente A1 y A2. Este predominio sugiere una tendencia a priorizar la accesibilidad y la claridad comunicativa, aunque puede limitar la riqueza léxica en contextos de mayor exigencia.

En los niveles intermedios (B1 y B2), los LLM muestran una presencia moderada de léxico más avanzado. En concreto, los porcentajes oscilan entre el 7,5 % y el 9,43 % en tareas de nivel B1, y entre el 7,23 % y el 11,5 % en tareas de nivel B2. Estos valores resultan comparables a los del corpus de referencia CAES, que registra un 8,43 % en B1 y un 7,23 % en B2, lo que indica un comportamiento relativamente alineado entre ambos en estas franjas de dificultad.

No obstante, las diferencias se hacen más evidentes en el nivel avanzado (C1). En este caso, los LLM presentan una menor proporción de vocabulario de alta complejidad, con valores que oscilan entre el 2,05 % y el 5,3 %, frente al 6,81 % observado en el CAES. Este desfase apunta a una menor capacidad de los modelos para incorporar de manera consistente léxico propio de niveles superiores, lo que podría afectar a la precisión y sofisticación de los textos generados en contextos académicos o especializados.

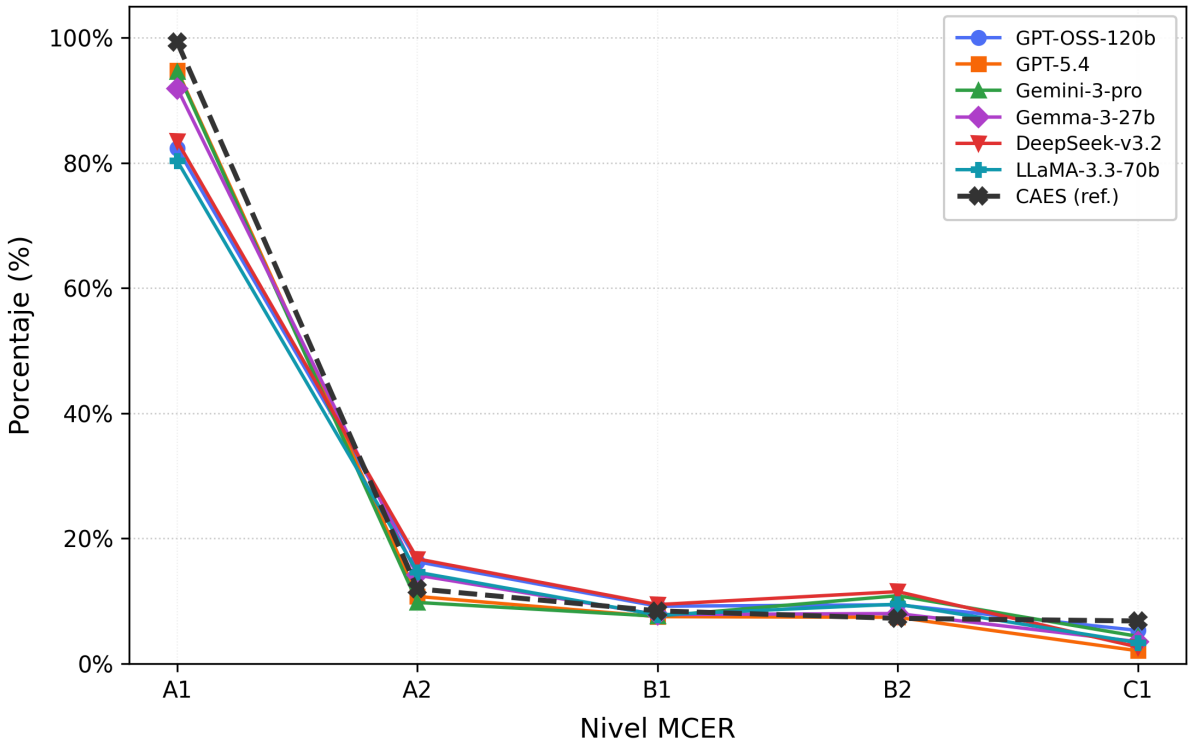

Figura 1: Gráfico con el % de léxico por niveles para cada uno de los modelos del estudio.

Estos resultados sugieren que, si bien los LLM reproducen de manera bastante fiel el uso léxico en niveles intermedios, todavía presentan limitaciones en la integración de vocabulario avanzado, donde los aprendices muestran una mayor especialización. En este sentido, la creciente proliferación de contenidos generados por IA, caracterizados por un léxico más estandarizado, podría tener implicaciones tanto en el aprendizaje de lenguas como en su evolución, favoreciendo dinámicas de simplificación o empobrecimiento léxico. En conjunto, estos hallazgos subrayan la necesidad de seguir profundizando en la evaluación y ajuste de los LLM en contextos educativos, con el fin de garantizar un uso pedagógico que promueva la riqueza y diversidad léxica.

Como líneas de trabajo futuro, se plantea profundizar en el análisis de la complejidad léxica de los textos generados mediante la incorporación de métricas específicas de riqueza, diversidad y sofisticación léxica. Asimismo, resulta de interés estudiar la influencia de la polisemia en los procesos de nivelación léxica, dado que una misma unidad léxica puede presentar distintos grados de complejidad en función del significado empleado y del contexto de uso. Estas líneas permitirán obtener una caracterización más precisa del comportamiento léxico de los modelos de lenguaje. Por último, nos proponemos también como trabajo futuro introducir en el *prompt* el nivel específico con el que se debe realizar la tarea, según su clasificación en el MCER, para comprobar si los modelos son capaces de generar un vocabulario adecuado al nivel indicado o si el léxico que utilizan es de niveles básicos, tal y como se ha observado en este estudio.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio para la Transformación Digital y de la Función Pública y por el Plan de Recuperación, Transformación y Resiliencia —financiado por el programa NextGenerationEU de la UE— en el marco del proyecto Desarrollo de modelos ALIA. También ha contado con el apoyo del Proyecto CONSENSO (PID2021-122263OB-C21), financiado por MCIN/AEI/10.13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR, del proyecto ROMANET (CERV-2024-CHAR-LITI -101215052), financiado por la Unión Europea en el marco del programa Ciudadanos, Igualdad, Derechos y Valo-

res, de los proyectos HEART-NLP-UJA (PID2024-156263OB-C21) y VERITAS-H (AIA2025-163322-C64), financiados por MICIU/AEI/10. 13039/501100011033 y por FEDER/UE, y del proyecto GALENO-IA (DGP_PIDI_2024_00852), financiado por la Unión Europea en el marco del Programa de la Estrategia de Especialización Inteligente para la Sostenibilidad de Andalucía (S4Andalucía) 2021-2027. El trabajo también ha sido financiado por la ayuda (FPI-PRE2022-105603) del Ministerio de Ciencia, Innovación y Universidades del Gobierno de España.

Bibliografía

Cantero Romero, M. V. 2025. El léxico ele en los modelos de lenguaje. *RILEX. Revista sobre investigaciones léxicas*, 8(1):155–185, Jan.

Conde, J., M. González, N. Melero, R. Ferrando, G. Martínez, E. Merino-Gómez, J. A. Hernández, y P. Reviriego. 2024. Open conversational llms do not know most spanish words. *Procesamiento del Lenguaje Natural*, 73(0):95–108.

Consejo de Europa. 2020. *Marco Común Europeo de Referencia para las Lenguas: Aprendizaje, Enseñanza, Evaluación. Volumen Complementario.* Servicio de Publicaciones del Consejo de Europa, Estrasburgo.

DeepSeek-AI. 2024. Deepseek-v3.

Gonzalez-Agirre, A., M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, I. Pikabea, A. Rubio, A. Shvets, A. Sallés, I. Lacunza, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, O. Pareras, V. Ruiz-Fernández, y M. Villegas. 2025. Salamandra technical report.

Google. 2025. Gemma 3 model overview.

Google. 2026. Models — gemini api.

Guo, Y., G. Shang, y C. Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.

Herbold, S., A. Hautli-Janisz, U. Heuer, Z. Kikteva, y A. Trautsch. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific Reports*, 13, 10.

Honnibal, M. y I. Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Instituto Cervantes. 2006. *Plan curricular del Instituto Cervantes. Niveles de referencia para el español.* Biblioteca Nueva, Madrid. 3 volúmenes.

Kendro, K., J. Maloney, y S. Jarvis. 2026. Do large language models produce texts with "human-like" lexical diversity? evidence from four chatgpt models. *International Journal of Applied Linguistics.* Accepted January 2026.

Kyle, K. y C. M. Berger. 2017. The tool for the automatic analysis of lexical sophistication (taales): version 2.0. *Behavior Research Methods*, 50, 07.

Laufer, B. y I. S. P. Nation. 1995. Vocabulary size and use: Lexical richness in l2 written production. *Applied Linguistics*, 16(3):307–322.

López-Mezquita Molina, M. T. 2007. La evaluación de la competencia léxica: test de vocabulario, su fiabilidad y validez.

Meta. 2024. Llama 3.3 model card and prompt formats.

Mullins, E., A. Portillo, K. Ruiz Rohena, y A. Piplai. 2024. Enhancing classroom teaching with llms and rag. En *Proceedings of the 25th Annual Conference on Information Technology Education*, SIGITE '24, página 145–146, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card.

OpenAI. 2026. Using gpt-5.4.

Palacios Martínez, I., F. M. Barcala Rodríguez, y G. Rojo. 2019. El corpus de aprendices de español (caes) y sus aplicaciones para la enseñanza y aprendizaje del español como lengua extranjera. En M. Blanco H. Olbertz, y V. Vázquez Rozas, editores, *Corpus y construcciones: Perspectivas hispánicas*,

volumen 79 de *Verba, Anexo*. Universidade de Santiago de Compostela, páginas 273–301.

Reviriego, P., J. Conde, E. Merino-Gómez, G. Martínez, y J. A. Hernández. 2024. Playing with words: Comparing the vocabulary and lexical diversity of chatgpt and humans. *Machine Learning with Applications*, 18:100602.

Skowron, D. y A. Baczkowska. 2023. Assessment of various gpt models versus the human text: a quantitative analysis of lexis and cohesion ocena różnych modeli gpt wzgledem tekstu napisanego przez człowieka: analiza leksyki i spójności tekstu. *Forum Filologiczne Ateneum*, 1, 10.

Universidad de Santiago de Compostela. 2022. Corpus de aprendices de español (caes), Marzo. Versión 2.1.